데이터마이닝응용

Term Project Final Report

지속가능한 전력 생산 - 전력 수요, 태양광 발전량, 풍력 발전량 예측

과목명	데이터마이닝응용
담당교수	이석룡 교수님
제출일	2023.12.16 (토)
학과	산업경영공학과
팀명	빵빵이들
학번/이름	이정현 201802798
	이의진 201902743 (PM)
	정경서 201903156
	최서여 202103464

목차

지속가능한 전력 생산

전력 수요, 태양광 발전량, 풍력 발전량 예측

I. 주제 선정 배경

1. 우리 삶에 필수적인 요소, 전기

전력은 4 차 산업혁명의 주춧돌이면서 문명사회의 근간을 이루는 필수 요소이다.
전력 수요를 잘못 예측하여 정부가 발전설비를 충분히 짓지 않게 되면 전국적으로 대정전을 겪을 수 있다. 이는 실제 2011 년 '9·15 순환 대정전'에서 신호 고장으로 도로를 아수라장으로 만들고, 기업들의 업무 마비 등과 같은 피해로 이어졌다. 그렇기 때문에, 첨단사회일수록 정전이 가져오는 피해도 커진다고 할 수 있고, 이 점이 바로 전력 수요를 제대로 예측하고 필요한 만큼만 발전소를 가동할 수 있도록 준비해야 하는 이유이다.

2. 지구 온난화의 주범, 화석 연료

최근 몇 십년 동안 온난화로 인한 기후 변화가 우리의 일상에 지속적인 영향을 미치고 있다. 이로 인해 화석 연료 사용량을 줄여야 할 필요성이 대두되고 있다. 신재생 에너지, 특히 태양광 및 풍력 에너지는 이러한 화석 연료 대체에 큰 기여를 할 수 있다.

3. 전력 수요 예측의 핵심 요인, 기후

전력 수요는 평균기온 등 기상상황과 밀접한 관련이 있다. 통상 평균기온이 오르면 에어컨 등 전력 다소비 기기 사용이 늘어 전력 수요가 급증하고, 평균 기온이 내리면 전력 수요가 줄어든다.

4. 전력 발전량 = 전력 수요량 – 친환경 발전량

전력 수요량과 친환경 발전량을 기상 상황을 바탕으로 예측하여, 필요한 전력발전량을 예측할 수 있다.

II. 프로젝트 목표

1. 전기 수요량, 풍력 발전량, 태양광 발전량 예측
2. 전기 수요량, 풍력, 태양광 발전량을 예측하여 화석 에너지를 이용한 전력 발전량을 도출

III. 데이터

구분	기간	범위	단위	변수
친환경 데이터 태양광, 풍력 발전량	2020 년부터 2022 년까지 3 개년	전국	시간	시간, 지역 발전량
전력데이터			시간	시간, 지역, 현재수요
기상데이터			5 분	시간, 지역, 기온, 강수, 기압, 습도, 풍향, 풍속

IV. 데이터 전처리 과정

데이터를 이해하고 데이터에 맞는 방법을 선택하여 분석하는 것과 그렇지 못한 분석의 결과는 크게 차이가 날 수 밖에 없다. 따라서 데이터를 분석하기에 앞서 데이터를 이해하고 그 특성을 잘 반영할 수 있도록 전처리하는 과정은 데이터 마이닝의 매우 중요한 부분이다. 다음은 분석에 앞선 데이터 전처리과정이다.

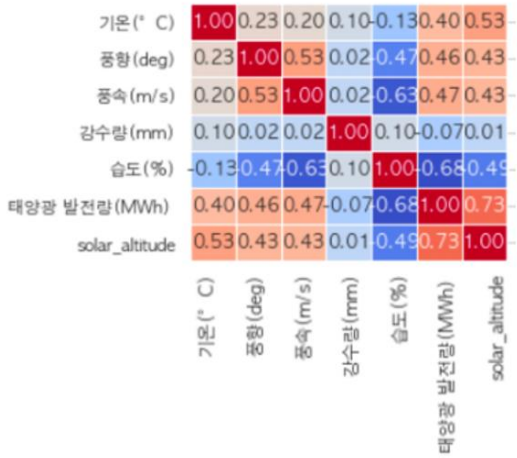
풍향

풍향의 경우 0-360 의 각도 값은 예측에 과도한 부하를 줄 뿐만 아니라 유의미한 차이라고 보기 어려워, 북, 북북동, 북동과 같이 16 방위로 구분하는 one-hot-encoding 을 활용하여 모델 예측에 도움을 주고자 한다.

태양 고도

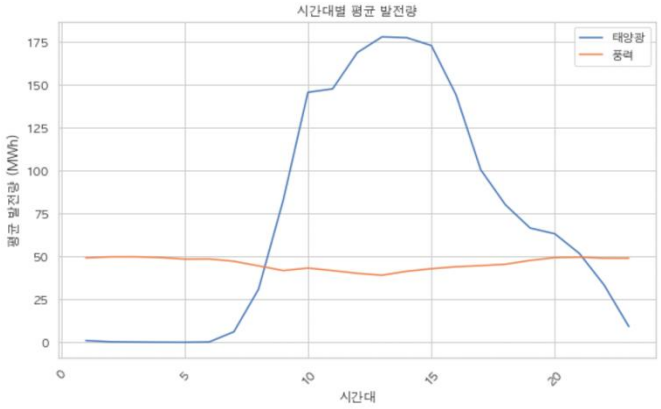
태양광 발전량과 유의미한 변수를 모색하던 중, 한 논문에서 태양의 고도가 큰 상관관계를 보인다는 내용을 접할 수 있었다. 태양의 고도를 도출하기 위해, 파이썬에서 제공하는 ‘pysolar’ 패키지를 활용해 각 지역별로 시간, 위도, 경도 정보를 입력하여 태양의 고도값을 출력 받았다. 이 값을 ‘태양 고도’ 변수로 사용하였다.

실제 태양 고도가 태양광 발전량에 유의미한 영향을 미치는지 확인하기 위해 피어슨 상관관계 분석을 진행하였다.



태양고도 – 태양광 발전량 상관관계 지수가 0.73 으로 높은 양의 상관관계를 가짐을 확인했다.

태양 발전 시간



다음은 태양광과 풍력 발전의 시간대별 발전량을 시각화한 그래프이다. 오전 8 시와 오후 8 시에 변곡점을 발견하고 태양광 발전에 대한 시간대를 오전 8 시부터 오후 8 시로 한정하였다.

지역

전력 데이터의 지점 구분은 16 개이다. 501 개의 지점으로 이루어진 친환경 데이터를 전력 데이터와 통합하기 위해 아래와 같이 지점을 통합하였다.

1. 16 개 지점의 중점 위도,경도 확보
2. 501 개 지점의 중점 위도,경도 확보
3. 16 개 지점의 위도,경도와 가장 가까운 기준으로 501 개의 지점을 16 개로 추출

이후 모델 예측에 사용할 데이터를 16 개의 지점의 변수를 설정하는 one-hot-encoding 을 진행하였다.

데이터 전처리

StandardScaler 는 scikit-learn 라이브러리에서 제공되는 데이터 전처리 도구 중 하나로, 데이터를 평균이 0 이고 표준 편차가 1 인 표준 정규 분포로 변환하는 역할을 한다, 데이터의 스케일이 서로 다를 때 사용하여 숫자 편차에 따른 데이터 중요도가 달라지는 경우를 방지한다

V. 기계 학습

Decision Tree (의사결정트리)

Decision Tree 는 데이터를 분석하고 이해하기 쉽도록 트리 구조로 표현된다. 루트 노드에서부터 시작하여 각 내부 노드에서는 정보이론 등을 통해 데이터를 분할하고, 각 가지로 이동하며 결정을 내려간다. 이 과정을 반복하여 잎(리프) 노드에 도달하면 최종적인 결정을 내린다. 해석이 쉽고 시각화가 용이하며, 범주형 및 수치형 데이터를 모두 다룰 수 있지만 적당한 멈춤 조건을 설정하지 않을 경우 과적합이 발생할 수 있다.

Random Forest (랜덤 포레스트)

Random Forest 는 여러 개의 결정 트리를 생성하고 그들의 예측을 결합하여 과적합을 줄이고 정확도를 향상시키는 앙상블(Ensemble) 학습 모델이다. 각 트리는 중복을 허용한 랜덤 샘플을 사용하여 독립적으로 학습하며, 각 노드에서는 랜덤하게 선택된 일부 속성으로 데이터를 분할한다. 예측 시에는 각 트리의 결과를 평균 또는 다수결로 결합한다. Random Forest 는 Decision Tree 에 비해 과적합을 줄이고 안정적인 예측을 제공하며, 다양한 유형의 데이터에 적용 가능하다. 하지만 Decision Tree 에 비하여 해석이 어려우며, 모델 구성에 따른 하이퍼파라미터 튜닝이 필요하다.

Multi-Layer Perceptron (다층 퍼셉트론, MLP)

MLP 는 인공 신경망의 한 종류로, 여러 개의 은닉층을 가진다. 입력층, 은닉층, 출력층으로 이루어져 있으며, 각 층은 연결된 노드(뉴런)들로 구성되어 있다. 입력층에서 시작하여 각 은닉층의 노드는 입력을 받아 가중치를 곱하고 활성화 함수를 통과시킨다. 이를 통해 신경망은 비선형성을 학습할 수 있다. 출력층에서는 최종 예측을 수행한다. 복잡한 비선형 패턴을 학습할

수 있으며, 다양한 유형의 데이터에 적용 가능하다. 하지만 대량의 데이터와 많은 연산이 필요하며, 과적합이 발생할 수 있다. 랜덤 포레스트와 마찬가지로 적절한 정규화 및 하이퍼파라미터 튜닝이 필요하다.

데이터 탐색 방법

Grid Search 는 머신 러닝 모델의 하이퍼파라미터 튜닝을 자동화하기 위한 탐색 방법이다. 모델의 성능을 최대화하기 위해 여러 가지 하이퍼파라미터 조합을 시스템적으로 탐색하는 방법으로, 가능한 모든 조합을 시도하여 최적의 조합을 찾는다 또한 Grid Search 에서는 교차 검증과 함께 사용됨으로 이 과정에서 모델을 여러 번 훈련하고 검증하여 성능을 평가한다. 이를 통해 모델이 데이터 셋에 과적합되는 것을 방지할 수 있다.

VI. 시계열분석

1. 시계열

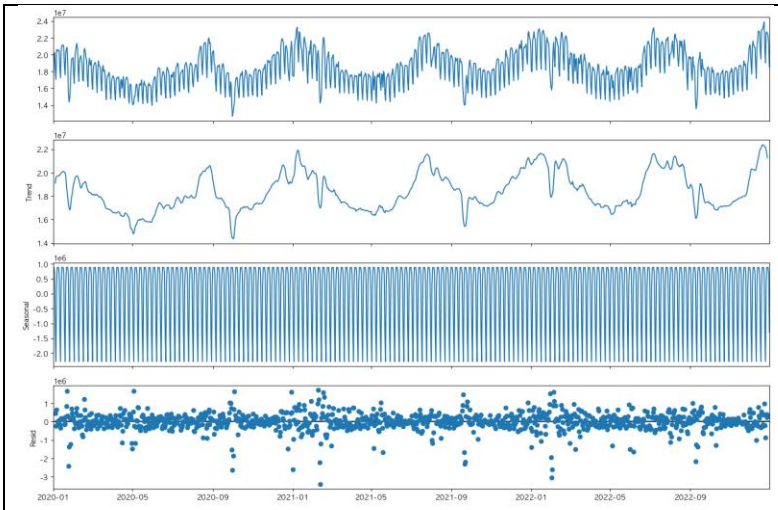
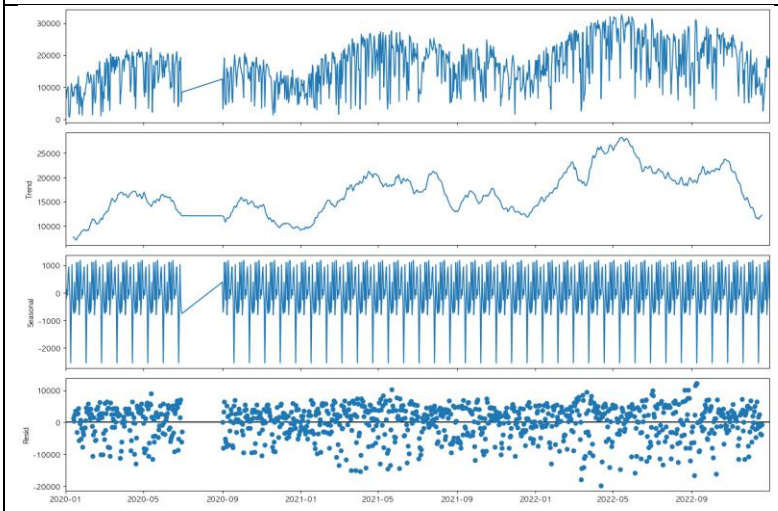
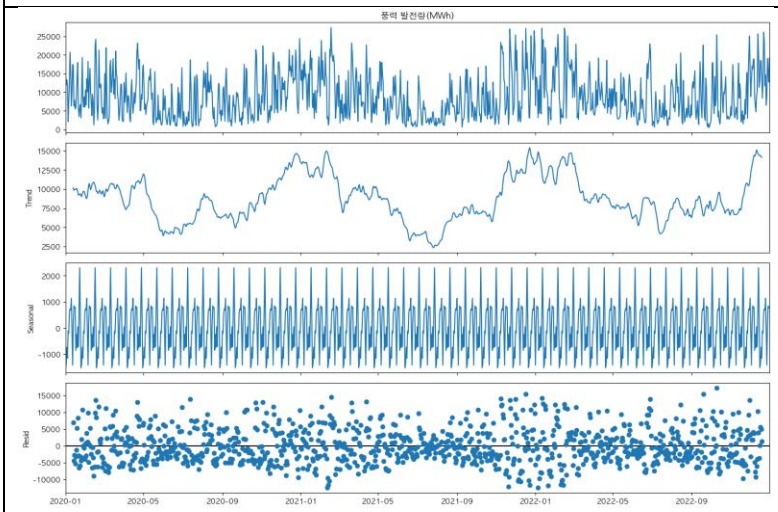
시계열 데이터란 시간에 대해 순차적으로 관측되는 데이터의 집합을 말한다. 쉽게 말해 데이터 분석 혹은 모델링에서 독립변수(independent variable)를 이용해서 종속변수(dependent variable)을 예측하는 방법이 일반적이라면, 시계열 데이터 분석은 시간을 독립변수로 활용한다고 생각하면 된다. 이러한 시계열 데이터를 활용함으로써 미래 시점의 데이터를 예측하거나, 일정한 길이의 시계열 데이터를 이용해서 패턴을 분류하는 분석을 진행할 수 있다.

본 프로젝트에서는 시간에 대해 순차적으로 관측된 전력 수요 데이터에 시계열분석을 적용하여 전력 수요량을 예측하고자 한다.

전통적으로 시계열 데이터 분석은 AR(Auto Regressive), MA(Moving Average), ARMA(Autoregressive Moving average), ARIMA(Autoregressive Integrated Moving average) 등과 같은 다양한 모델을 통해 불규칙적인 시계열 데이터에 규칙성을 부여하는 방식을 활용해왔다.

다양한 시계열분석 방법 중 전력 수요 데이터에 적용할 시계열분석 방법을 선택하기 위해 아래와 같이 시계열분해를 진행하였다. 시계열분해는 시계열 데이터를 구성하는 다양한 구성 요소를 분리하여 각 구성 요소의 특성을 더 잘 이해하고 모델링하기 위한 기술을 의미한다.

- 시계열 분해 결과

	1) 전력 수요 데이터 특징 (1) 단기 예측 (2) 추세(Trend) 존재 X (3) 데이터(Sample)(표본)의 크기 > 50
	2) 태양광 데이터 특징 (1) 단기 예측 (2) 추세(Trend) 존재 X (3) 데이터(Sample)(표본)의 크기 > 50
	3) 풍력 데이터 특징 (1) 단기 예측 (2) 추세(Trend) 존재 X (3) 데이터(Sample)(표본)의 크기 > 50

- Observed (관측값): 원본 데이터 열
- Trend (추세): 데이터의 장기적인 변동(증가, 감소)을 보여주는 부분

- Seasonal (계절성): 주기적으로 반복되는 패턴을 보여주는 부분
- Residual (잔차): 추세와 계절성을 제외한 나머지 부분으로, 모델이 설명하지 못하는 남은 변동이 표시된다.

시계열 분해를 진행한 결과, 전력, 태양광, 풍력 데이터가 위와 같은 3 가지 조건을 만족하기에 ARIMA 시계열분석을 적용하였다.

2. ARIMA 시계열분석

가) 정상성 검정

많은 통계적 시계열 모델들(AR, MA, ARMA, ARIMA 등)은 시계열 데이터의 정상성을 가정하고 있다. 본 프로젝트에서는 선택한 ARIMA 시계열분석을 적용하기 전, 시계열 데이터의 필수 조건인 정상성(Stationary)을 파악하고자 한다.

정상성이란 '데이터 변동의 안정성'을 의미하며, 정상성(Stationary)을 가진 데이터란 일관된 평균, 분산(표준편차) 및 공분산, 자기 상관(auto-correlation)정도를 보이는 데이터를 말한다. 요약하자면, 정상 시계열은 시간의 흐름에 따라 그 통계적 특성이 변하지 않는 시계열을 의미한다. 이를 파악하기 위한 검정 방법으로는 KPSS(Kwiatkowski-Phillips-Schmidt-Shin)검정과 ADF(Augmented Dicky-Fuller)검정이 있다. 일반적으로 정상성 검정에 있어서는 1 가지 검정만을 사용하지 않고, 다양한 검정 방법을 사용하기에 3 가지 데이터의 정상성 검정에 있어서 위 2 가지 방법을 이용하고자 한다.

1) Kwiatkowski-Phillips-Schmidt-Shin(KPSS) 검정

KPSS 검정은 일종의 단위근 검정으로 95%의 신뢰도를 바탕으로 0.05 유의수준을 두었을 때 아래와 같은 가설 설정을 기반으로 한다. Python 에서는 statsmodels 라이브러리에서 kpss 를 import 하여 사용할 수 있다.

- H0: 해당 시계열은 정상 시계열이다.
- H1: 해당 시계열은 비정상 시계열이다.

p-value <= 0.05 이면 H0 기각, H1 채택으로 비정상 시계열 데이터이다.
p-value > 0.05 이면 H0 채택, H1 기각으로 정상 시계열 데이터이다.

2) Augmented Dickey-Fuller(ADF) 검정

ADF 검정은 단위근(Unit-root)검정으로 95%의 신뢰도를 바탕으로 0.05 유의수준을 두었을 때 아래와 같은 가설 설정을 기반으로 한다. Python 에서 statsmodels 라이브러리에서 adfuller 를 import 하여 검정 적용이 가능이 가능하다. 이는 KPSS 검정과 함께 자주 활용된다.

- H0: 해당 시계열은 비정상 시계열이다. (시계열에 단위근이 존재한다.)
- H1: 해당 시계열은 정상 시계열이다.

p-value <= 0.05 이면 H0 기각, H1 채택으로 정상 시계열 데이터이다.

p-value > 0.05 이면 H0 채택, H1 기각으로 비정상 시계열 데이터이다.

KPSS 검정과 귀무가설이 반대이다.

KPSS 검정 코드

```
# 1. KPSS 검정

kpss_test_result = kpss(ts)

print(f'KPSS Statistic: {kpss_test_result[0]}')
print(f'p-value: {kpss_test_result[1]}')
print(f'Lags Used: {kpss_test_result[2]}')
print(f'Critical Values: {kpss_test_result[3]}')

print(f'검증결과: {"비정상(non-stationary)" if kpss_test_result[1] <= 0.05 else "정상(stationary)"} 시계열 데이터입니다.')
```

ADF 검정 코드

```
# 2. ADF 검정

adf_test_result = adfuller(ts, autolag='AIC')

print(f'ADF Statistic: {adf_test_result[0]}')
print(f'p-value: {adf_test_result[1]}')
print(f'Critical Values: {adf_test_result[4]}')
print(f'Best number of lags (AIC): {adf_test_result[3]}')

print(f'검증결과: {"정상(stationary)" if adf_test_result[1] <= 0.05 else "비정상(non-stationary)"} 시계열 데이터입니다.')
```

1. 전력 데이터

```
KPSS Statistic: 0.7430466391108305
p-value: 0.01
Lags Used: 16
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
검증결과: 비정상(non-stationary) 시계열 데이터입니다.
```

KPSS 검정 결과 p-value 0.01<=0.05 로, 비정상 시계열 데이터이다. 비정상 시계열 데이터를 정상 시계열 데이터로 바꾸어 주기 위해 차분을 진행한다.

```
ADF Statistic: -3.3802295339986768
p-value: 0.011651595050739783
Critical Values: {'1%': -3.436459052172655, '5%': -2.864237372528562, '10%': -2.568206176974609}
Best number of lags (AIC): 1073
검증결과: 정상(stationary) 시계열 데이터입니다.
```

ADF 검정 결과 p-value 0.01<=0.05 로, 정상 시계열 데이터이다.

2. 태양광 데이터

```
KPSS Statistic: 2.2632948975972504
p-value: 0.01
Lags Used: 17
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
검증결과: 비정상(non-stationary) 시계열 데이터입니다.
```

KPSS 검정 결과 p-value 0.01<=0.05 로, 비정상 시계열 데이터이다.

```
ADF Statistic: -3.659369369268652
p-value: 0.004722346360779797
Critical Values: {'1%': -3.4367709764382024, '5%': -2.8643749513463637, '10%': -2.568279452717228}
Best number of lags (AIC): 1021
검증결과: 정상(stationary) 시계열 데이터입니다.
```

ADF 검정 결과 p-value 0.005<=0.05 로, 정상 시계열 데이터이다.

3. 풍력 데이터

```
KPSS Statistic: 0.15591509243174848
p-value: 0.1
Lags Used: 14
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
검증결과: 정상(stationary) 시계열 데이터입니다.
```

KPSS 검정 결과 p-value 0.1>=0.05 로, 정상 시계열 데이터이다.

```
ADF Statistic: -4.065860472999209
p-value: 0.0011023148814966494
Critical Values: {'1%': -3.436425000208065, '5%': -2.864222352544219, '10%': -2.5681981773275466}
Best number of lags (AIC): 1079
검증결과: 정상(stationary) 시계열 데이터입니다.
```

ADF 검정 결과 p-value 0.001<=0.05 로, 정상 시계열 데이터이다.

나) 차분

시계열분석을 진행할 때에는 일반적으로 로그변환 및 차분을 통해 정상성을 확보한 후에 자기회귀 모형을 구축하는 것이 특징이다. 정상성을 나타내지 않는 데이터는 복잡한 패턴을 모델링하여 분석하기 어렵기 때문이다.

더불어 정상성을 가진 데이터로 만드는 이유는 시간의 흐름에 따라 증가 혹은 감소 추세가 있는 현상 또는 주기적인 증감이 있는 현상을 보이는 이슈를 연구할 때, 자연 발생적이거나 문제의 예측변수와 관계 없는 요인들의 영향력을 배제하고 순수한 예측변수의 힘을 파악하고자 하기 때문이다.

가) 정상성 검정에서의 결과, KPSS 검정에서 전력 데이터와 태양광 데이터가 비정상 시계열 데이터라는 결과를 얻었으므로 정상성을 만족시키기 위해 차분을 진행하고자 한다.

차분이란 이어진 데이터들의 차이를 구하는 것을 말하며 1 번의 차이를 구하는 것을 1 차 차분, 1 차 차분값을 다시 차분하는 것을 2 차 차분이라 한다. 이와 같이 차분은 데이터의 길이가 충분할 경우 여러 번 수행될 수 있다. (하지만 대부분의 경우, 1 차 차분만으로 정상적인 시계열이 만들어지며, 2 차 이상 차분을 할 경우 해당 데이터에 적합한 모델의 설명력이 낮아지며 데이터의 소실이 커진다.)
차분을 수식으로 나타내면 아래와 같다. (t)시점의 값에서 (t-1)시점의 값을 빼는 것이다.

$$y_t^* = y_t - y_{t-1}$$

차분을 통해 시계열의 수준에서 나타나는 변화를 제거하면 시계열의 평균 변화를 일정하게 만드는데 도움이 된다. 결과적으로 추세나 계절성이 제거 혹은 감소된다.
아래는 차분을 진행한 결과이다.

정상성 확보를 위한 차분

```
# 1차 차분

ts_diff = ts.diff().dropna()
# 차분을 진행할 때는 두 값을 빼주게 됨. 그렇기에 차분을 진행할 때마다 데이터의 길이가 짧이지면서, 짧아진 부분에 공백이 생김.
# 이 공백을 지워주지 않으면 ADF검정 과정에서 오류가 발생할 수 있다고 함.
```

아래는 1 차 차분을 진행한 데이터를 바탕으로 KPSS 및 ADF 검정을 진행한 결과이다.

1. 전력 데이터

```
KPSS Statistic after differencing: 0.036706817180015706
p-value after differencing: 0.1
Lags Used after differencing: 89
Critical Values after differencing: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
검증결과: 정상(stationary) 시계열 데이터입니다.
```

KPSS 검정 결과, 1 차 차분한 전력 데이터는 p-value 0.1 >= 0.05 로, 정상 시계열 데이터이다.

```
ADF Statistic: -7.36492581928262
p-value: 9.29014868301968e-11
Critical Values: {'1%': -3.4364533503600962, '5%': -2.864234857527328, '10%': -2.568204837482531}
Best number of lags (AIC): 1074
검증결과: 정상(stationary) 시계열 데이터입니다.
```

ADF 검정 결과, 1 차 차분한 전력 데이터는 p-value <= 0.05 로, 정상 시계열 데이터이다.

2. 태양광 데이터

```
KPSS Statistic after differencing: 0.07086770191039689
p-value after differencing: 0.1
Lags Used after differencing: 123
Critical Values after differencing: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
검증결과: 정상(stationary) 시계열 데이터입니다.
```

KPSS 검정 결과, 1 차 차분한 태양광 데이터는 p-value 0.1 >= 0.05 로, 정상 시계열 데이터이다.

```
ADF Statistic: -15.714542205094503
p-value: 1.346828176793479e-28
Critical Values: {'1%': -3.4367709764382024, '5%': -2.8643749513463637, '10%': -2.568279452717228}
Best number of lags (AIC): 1021
검증결과: 정상(stationary) 시계열 데이터입니다.
```

ADF 검정 결과, 1 차 차분한 태양광 데이터는 p-value 1.35 >= 0.05 로, 정상 시계열 데이터이다.

다) ARIMA 시계열분석 진행 및 결과

ARIMA 모델은 자기 회귀(AR) 차수, 차분(Differencing) 차수, 이동 평균(MA) 차수의 세 가지 구성 요소로 이루어져 있다. 여러 매개변수 조합에 대해 모델을 훈련하고 AIC(Akaike Information Criterion)를 통해 모델의 성능을 평가할 수 있다. AIC 는 모델이 데이터에 얼마나 잘 맞는지를 측정하는 지표로서, 모델의 복잡성과 데이터 적합도를 고려하여 최적 모델을 선택하는 데 사용된다. AIC 가 낮을수록 모델의 성능이 좋다고 판단한다.

ARIMA 모델을 사용하여 전력 시계열 데이터를 예측하는 과정을 진행하였다. 먼저, 전체 데이터셋을 학습 데이터와 검증 데이터로 분리하고, 가능한 다양한 매개변수 조합을 통해 Grid Search 를 수행하여 최적의 ARIMA 모델을 찾았다.

Grid Search 를 통해 찾아진 최적 매개변수를 이용하여 전체 데이터셋에 ARIMA 모델을 다시 훈련하고, 학습 데이터 이후의 구간에 대해 예측을 수행하였다. 이후 예측 결과와 실제 검증 데이터를 비교하여 모델의 성능을 평가하였으며, 이때 평가 지표로는 Mean Squared Error(MSE), Root Mean Squared Error(RMSE), 그리고 Mean Absolute Percentage Error(MAPE)가 사용되었다.

1. 전력 데이터

```
Best ARIMA Order (by AIC): (1, 1, 1)
Best AIC: 27183.299496020853
Mean Squared Error on Validation Data: 1839853765927.0137
Root Mean Squared Error on Validation Data: 1356412.0929595893
Mean Absolute Percentage Error on Validation Data: 0.0518
```

2. 태양광 데이터

```
Best ARIMA Order (by AIC): (1, 1, 1)
Best AIC: 16527.18767516721
Mean Squared Error on Validation Data: 31893916.24652529
Root Mean Squared Error on Validation Data: 5647.46989779718
Mean Absolute Percentage Error on Validation Data: 0.3493
```

3. 풍력 데이터

```
Best ARIMA Order (by AIC): (1, 0, 1)
Best AIC: 17405.603062277896
Mean Squared Error on Validation Data: 23637708.764572266
Root Mean Squared Error on Validation Data: 4861.862684668528
Mean Absolute Percentage Error on Validation Data: 0.8908
```

VI 결과 비교

일단위

데이터	측정	시계열	DT	RF	MLP
태양광	RMSE	5600	5284	4507	4521
	MAPE	0.349	0.5	0.376	0.329
풍력	RMSE	5000	3539	2713	2713
	MAPE	0.8	0.49	0.37	0.35
전력	RMSE	135 만	169 만	162 만	162 만
	MAPE	0.05	0.0747	0.070	0.071

시간단위

데이터	측정	DT	RF	MLP
태양광	RMSE	70	54	55
풍력	RMSE	44	40	60

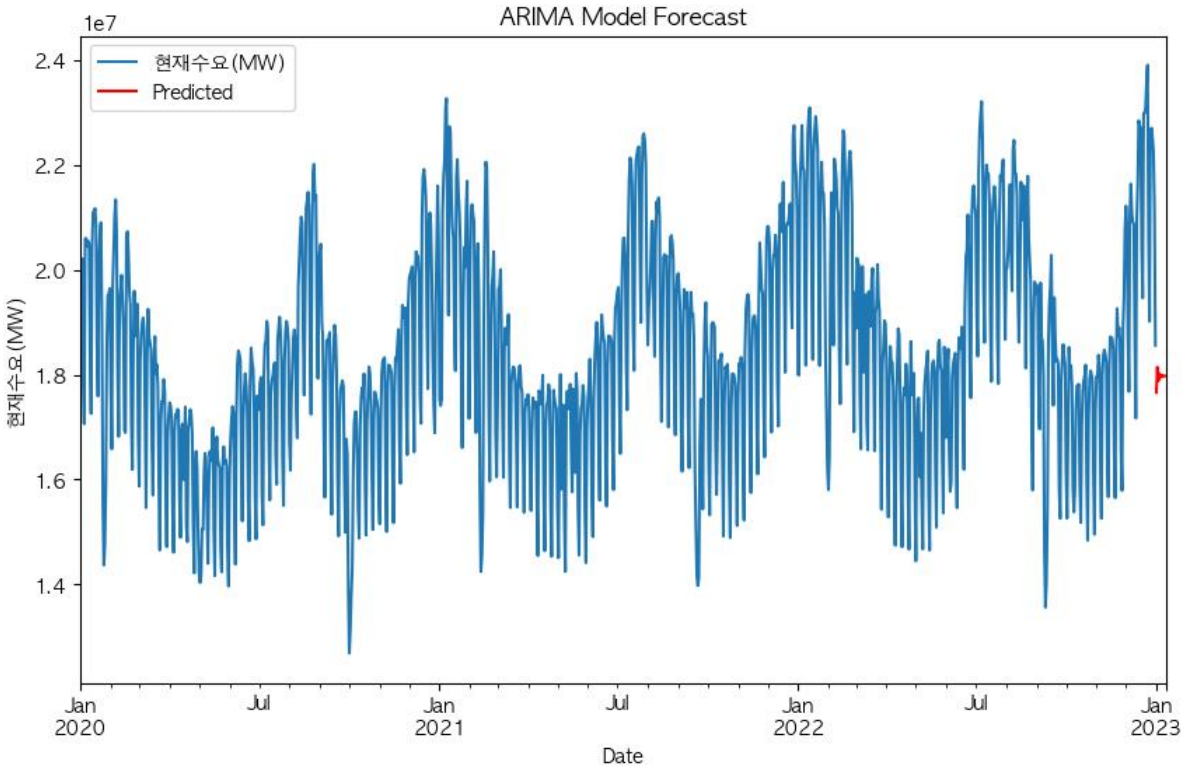
VII. 최종 결론

최종적으로 태양광,풍력에서는 Random Forest, 전력에서는 시계열 분석이 가장 좋은 지표를 보여줌에 따라서 해당 모델을 선정하였다.

2023 년 1 월 1 일부터 10 일간 시계열분석 예측 결과

```
2023-01-01    1.766077e+07
2023-01-02    1.814404e+07
2023-01-03    1.788392e+07
2023-01-04    1.802393e+07
2023-01-05    1.794857e+07
2023-01-06    1.798913e+07
2023-01-07    1.796730e+07
2023-01-08    1.797905e+07
2023-01-09    1.797273e+07
2023-01-10    1.797613e+07
2023-01-11    1.797430e+07
```

예측 결과에 대한 시각화



다음은 해당 모델을 바탕으로 예측한 값이다.

2023.01.01 (월) 예측

총 전기 수요량	17,660,770	(실제 : 17,248,132.39)
풍력 발전 예측량	16,064	(실제 : 19,403.73)
태양광	7,061	(실제 : 19,052.48)
화석 연료	17,637,645	(실제 : 17,209,676.2)

단위 : MWh

태양광이 실제 값보다 매우 크게 오차가 나타난 것의 이유로는 태양이 뜨고 지는 시점을 계절마다 고려했어야 하는데 연평균의 시간단위당 발전량을 바탕으로 고려하여 계절별로 편차가 크게 나타나 8 시부터 20 시까지의 일조량이 여름보다 작은 1 월의 값이 작게 나온 것으로 보고 있다.

한계점

먼저 모든 신재생 에너지 발전량을 고려한 것이 아니기 때문에 실제 화석연료 발전이 필요한 양을 예측했다고 보기에는 어렵다는 한계점이 있었다. 때문에, 다양한 에너지원을 바탕으로 전력을 생산하는 현실을 추가 고려할 필요가 있다고 생각한다.

분석에서 Grid Search 를 통해 탐색한 하이퍼 매개변수들은 Random Forest 의 깊이와 MLP 의 max_iter (max iteration) 밖에 없었다. 따라서 높은 정확성을 위해서는 다양한 하이퍼 매개변수에 대해서 탐색을 진행해야 한다.

전력 수요 시계열의 경우, 전력 수요가 높은 시간대를 따로 분리해서 분석하면 더 좋은 결과를 얻을 수 있을 것이라 기대되기 때문에 이를 보완점으로 생각한다.

풍력/태양광 발전 데이터 분석에서 앙상블 기법으로 Voting Regressor 를 사용해 보았지만, 더 좋은 결과를 얻을 수 없었다. 분석에서의 앙상블에 대한 이해가 부족하기 때문이라고 생각했고, 그렇기에 다양한 기법을 수행하지 못한 점을 한계점으로 생각할 수 있다.

VIII. 견해

이의진(PM)

프로젝트를 진행할 때 제일 난항을 겪었던 부분은 주제를 선정하는 일이었습니다. 어떤 주제를 선정해야 의미 있는 결과를 도출할 수 있을까? 어떤 데이터를 고려해야 할까 많은 고민들을 팀원들과 함께 고민해 보았습니다. 이 과정에서 많은 생각들이 오고 갔고 이 프로젝트 중에 가장 많이 의견을 교환했던 부분이 바로 이 때였던 것 같습니다. 이를 통해 주제 선정 단계에서 팀원들 간의 의견 교환과 토의를 통해 어떤 문제를 해결하고 어떤 방향으로 나아갈지에 대한 합의를 이루는 것이 향후 프로젝트 전반에 큰 영향을 미친다는 것을 배웠습니다.

프로젝트를 진행하면서 새로운 방법을 찾고 적용하는 부분에서도 많은 배움이 있었지만, 프로젝트에서 가장 많이 배웠던 점은 팀원들 과의 소통을 통해 결과를 도출하고 각자의 일을 정하고 스케줄링 하는 등 팀 프로젝트를 진행하는 것이었습니다. 중간에 서로의 조금 다른 생각이 최종 결과에서는 어떻게까지 엇갈릴 수 있는지 알 수 있었고, 프로젝트의 진행을 위해서는 끊임없는 소통이 제일 중요한 것을 느꼈습니다.

혼자서는 진행하지 못하였을 것 같은 부분이 서로의 채움을 통해 완성되는 프로젝트를 진행하면서, 다음 팀프로젝트에서는 소통과 협업을 통해 보다 공동의 프로젝트를 진행할 수 있는 점을 배웠습니다.

정경서

전체 프로젝트를 통틀어, 지속 가능한 전력 생산을 목표로 친환경 데이터를 활용하여 화력 발전 에너지양을 줄이기 위한 노력을 기울였습니다. 이는 화력 발전을 최소화하여 환경에 미치는 영향을 최소화하고 지속 가능한 에너지 소스를 촉진하는 데에 기여하는 중요한 목표였습니다. 따라서 프로젝트를 통해 지속 가능한 에너지 사용에 대한 의의를 깊이 이해하고 실제 데이터 예측을 통해 이를 구현하는 경험을 쌓을 수 있었습니다.

본인이 주로 중점을 둔 시계열 데이터 분석 과정에서는 수업에서 다루지 않은 시계열 분석을 개인적으로 공부하며 데이터 분석에 대한 심층적인 이해를 높일 수 있는 소중한 기회가 되었습니다. 시계열 분석의 다양한 종류와 실제 데이터에 적용되는 예시, 그리고 데이터 분석을 위한 전제 조건 등에 대해 학습하면서 새로운 지식을 습득할 수 있었습니다. 특히 ARIMA 시계열 분석을 활용하여 전력 수요 데이터를 예측하는 과정은 흥미로운 도전이었고, 통계적인

측면에서도 다양한 인사이트를 얻을 수 있었습니다. 이러한 경험을 통해 지속 가능한 에너지 사용과 데이터 분석의 결합이 현실적인 문제 해결에 어떻게 기여할 수 있는지에 대한 실질적인 통찰력을 얻을 수 있었습니다.

이러한 배움을 얻을 수 있었던 건 팀원 간의 원활한 소통 덕분이었다고 생각합니다. 끊임없이 소통하고 방향성을 맞추며 프로젝트를 진행한 덕분에 서로의 의견을 오판하는 상황이 적게 발생했고, 그 덕에 프로젝트가 원활하고 매끄럽게 진행되었다고 생각합니다. 또한, PM 의 편향되지 않는 업무 분배 및 각 팀원의 배경지식을 바탕으로 이루어진 효율적인 업무 분배가 무엇보다 원활한 팀 프로젝트를 이끌었다고 생각했고, PM 역할의 중요성을 한번 더 깨닫게 되는 경험을 얻을 수 있었습니다.

최서여

프로젝트를 진행하면서 수업에서 배운 내용 외에도 새로운 도전이 있었습니다. 특히, 시계열 분석이라는 주제는 처음 접하는 분야였습니다. 초기에는 낯선 용어와 개념 때문에 어려움을 겪었지만, 이를 극복하고자 노력하는 과정에서 많은 것을 배우게 되었습니다.
우선, 시계열 데이터의 특성과 패턴을 이해하는 것이 중요하다는 것을 깨닫게 되었습니다. 데이터의 순서와 시간 간격에 대한 고려가 예측 모델을 개발하는 데에 큰 영향을 미치는 것을 이해했습니다. 이를 통해 예전에는 간과하던 세부 사항들이 실제로 모델의 성능에 큰 영향을 미칠 수 있음을 깨달았습니다.
또한, 다양한 시계열 분석 기법을 학습하고 적용해보는 경험은 이론을 현실에 적용하는 데 도움이 되었습니다. 모델의 선택, 파라미터 튜닝, 그리고 결과 해석에 대한 과정에서 실제적인 문제 해결 능력이 향상되었습니다.

프로젝트를 통해 새로운 도구와 기술을 배우는 것은 어려움이 따르기 마련입니다. 그러나 이러한 도전을 통해 더욱 폭넓은 시야를 가질 수 있었고, 이는 졸업 프로젝트에서도 큰 도움이 될 것이라고 믿습니다.

이정현

데이터를 수집하고 전처리하는 과정에서 다양한 기법과 많은 노력이 필요함을 느꼈습니다. 특히 같은 데이터라도, 어떤 방식으로 전처리하여 모델 예측을 진행하는지에 따라 다른 결과가 나올 수 있고, 그 정도가 무시하지 못할 정도로 중요하다는 것을 직접 확인해보니 놀라웠습니다.

선행된 수많은 학술 연구들을 바탕으로, 우리 데이터의 특성에 맞는 여러가지 변수를 연관성 분석으로 검토하여 추가 채택, 변환하는 등의 과정으로 데이터의 질을 높여 보다 높은 예측 정확도를 가져가는 과정 또한 매력적으로 느껴졌습니다.

데이터마다 예측에 필요한 학습법을 한정하지 않고 여러 후보군 중 가장 좋은 정확도를 가진 모델을 채택하는 방식이 참 맘에 들었습니다. 데이터의 특성마다 다른 기법을 적절히 활용하여 우수한 예측치를 도출하고, 이를 바탕으로 하나의 결론을 도출하는 과정이 매력적이었습니다.

지속 가능한 가치를 창출한다는 주제에 맞게, 예측한 결과값을 좀 더 보완한다면 충분히 화석연료를 이용한 발전량을 실제 전력 수요값에 근접한 수준으로 예측하여, 보다 효율적이고 경제적이며 친환경적인 의사결정을 내린다는 부분이 상당히 유의미한 것 같습니다.

이번 데이터 마이닝 팀 프로젝트를 통해 충분한 가치를 지닌 결론을 도출할 수 있었습니다. 이런 양질의 경험을 할 수 있게 많은 시간 토의하고 연구했던 팀원들에게 고마움을 전하고 싶습니다.

IX. 참고 문헌

송경빈, 문찬호 and 권보성. (2022). 한국 전력시스템의 240 시간 전력수요예측에 대한 딥러닝 모델과 학습기법. 전기학회논문지, 71(4), 585-591.

김지은, 천관호. (개최날짜). LSTM 을 활용한 단기 전력수요 예측기법. 대한전기학회 학술대회 논문집, 개최지.

권보성, 전재성, 공병철. (2023). 단기 전력수요예측이 한국의 전력시장 가격에 미치는 영향 분석. 대한전기학회 학술대회 논문집, 개최지.

차현종, 강아름. (2023). 기상정보를 활용한 머신러닝 기반의 전력수요 예측 모델. 한국콘텐츠학회논문지, 23(2), 117-124, 10.5392/JKCA.2023.23.02.117

김형욱,「날씨 따라 바뀌는 전력수요 예측…한전, 기상청 빅데이터 공유 확대」,『이데일리』 2022.07.12, https://m.edaily.co.kr/news/Read?newsId=03289846632393864&mediaCodeNo=257&utm_source=https://www.google.com/(2023-11-03 접속).

부형권,「심야전력 수요예측 잘못 한전 125 억 경영손실」.『동아일보』 2009-09-21, https://www.donga.com/news//article/all/20010204/7643869/1 (2023-11-03 접속).

성수영, 「전력수요 예측은 왜 번번이 틀리나」, 『생글생글』 2018.08.20, https://sgsg.hankyung.com/article/2018081702641(2023-11-03 접속).

편집팀,「[에릭인사이트] 4 차 산업혁명에는 '전력'이 핵심이다」,『전기신문』 2021.06.03, https://www.electimes.com/news/articleView.html?idxno=218285 (2023-11-06 접속).